Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package

Sonia Tarazona[1,2], Pedro Furió-Tarí[1], David Turrà[3],
Antonio Di Pietro[3], María José Nueda[4], Alberto Ferrer[2], and Ana Conesa[1,5]

June 9, 2015

[1]Genomics of Gene Expression Lab, Centro de Investigación Príncipe Felipe, Eduardo Primo Yúfera 3, 46012, Valencia, Spain

[2]Department of Applied Statistics, Operations Research and Quality, Universidad Politécnica de Valencia, Camí de Vera, 46022, Valencia, Spain

[3]Department of Genetics, Universidad de Córdoba, Campus de Rabanales Edificio Gregor Mendel, 14071, Córdoba, Spain

[4]Statistics and Operational Research Department, Universidad de Alicante, Carretera San Vicente del Raspeig s/n, 03690, Alicante, Spain

[5]Microbiology and Cell Science Department, Institute for Food and Agricultural Sciences, University of Florida, Gainesville, FL 32603, USA

Supplementary Material

Contents

1 About data sets used in this work

1.1 ENCODE samples

All the ENCODE samples were downloaded from ENCODE website:
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq/

These are the names of the fastq files for each of the samples we used in the analysis (paired-end sequencing):

Bcells (CD20)

Harbor_Cd20_Pap_Rp2

- wgEncodeCshlLongRnaSeqCd20CellPapFastqRd1Rep2.fastq.tgz
- wgEncodeCshlLongRnaSeqCd20CellPapFastqRd2Rep2.fastq.tgz

Harbor_Cd20_Pam_Rp1-2
For this condition, we summed the counts from replicates 1 and 2 in order to get comparable sequencing depth for both protocols (Pap and Pam):

- wgEncodeCshlLongRnaSeqCd20CellPamFastqRd1Rep1.fastq.tgz
- wgEncodeCshlLongRnaSeqCd20CellPamFastqRd2Rep1.fastq.tgz
- wgEncodeCshlLongRnaSeqCd20CellPamFastqRd1Rep2.fastq.tgz
- wgEncodeCshlLongRnaSeqCd20CellPamFastqRd2Rep2.fastq.tgz

Monocytes: CD14

Harbor_Monocd14_Pap_Rp2

- wgEncodeCshlLongRnaSeqMonocd14CellPapFastqRd1Rep2.fastq.tgz
- wgEncodeCshlLongRnaSeqMonocd14CellPapFastqRd2Rep2.fastq.tgz

Harbor_Monocd14_Pam_Rp1-2
For this condition, we summed the counts from replicates 1 and 2 in order to get comparable sequencing depth for both protocols (Pap and Pam):

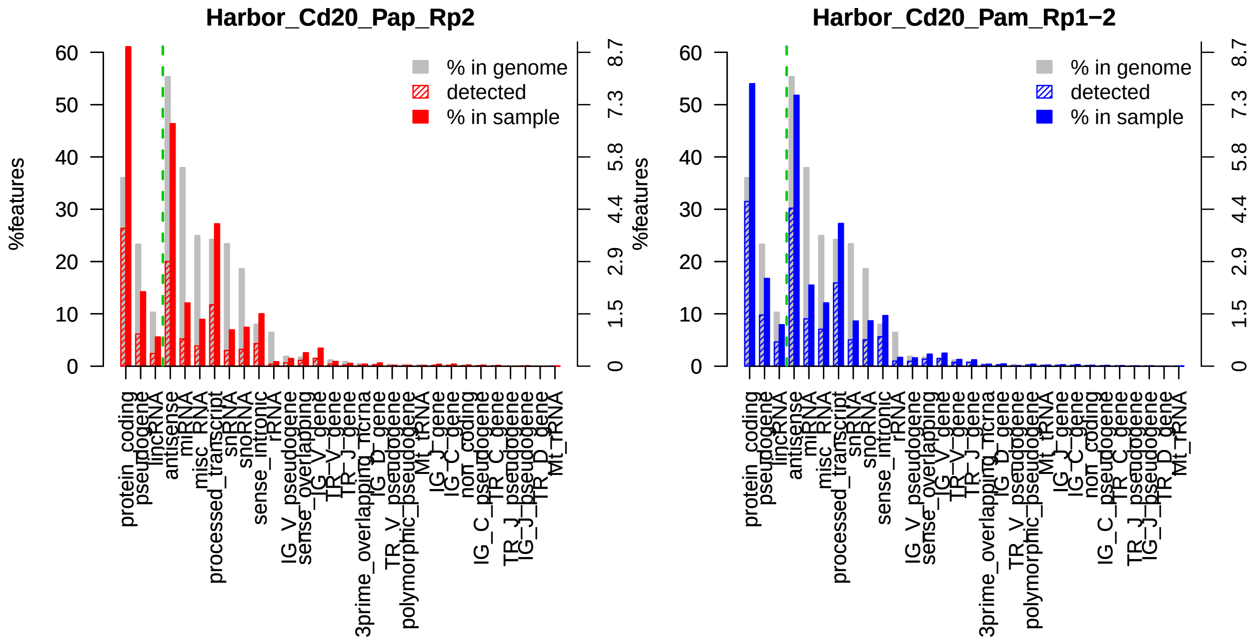
- wgEncodeCshlLongRnaSeq/wgEncodeCshlLongRnaSeqMonocd14CellPamFastqRd1Rep1.fastq.tgz
- wgEncodeCshlLongRnaSeq/wgEncodeCshlLongRnaSeqMonocd14CellPamFastqRd2Rep1.fastq.tgz
- wgEncodeCshlLongRnaSeq/wgEncodeCshlLongRnaSeqMonocd14CellPamFastqRd1Rep2.fastq.tgz
- wgEncodeCshlLongRnaSeq/wgEncodeCshlLongRnaSeqMonocd14CellPamFastqRd2Rep2.fastq.tgz

2 Quality control of expression data

2.1 Biotypes.

2.1.1 “Biotype detection” plots

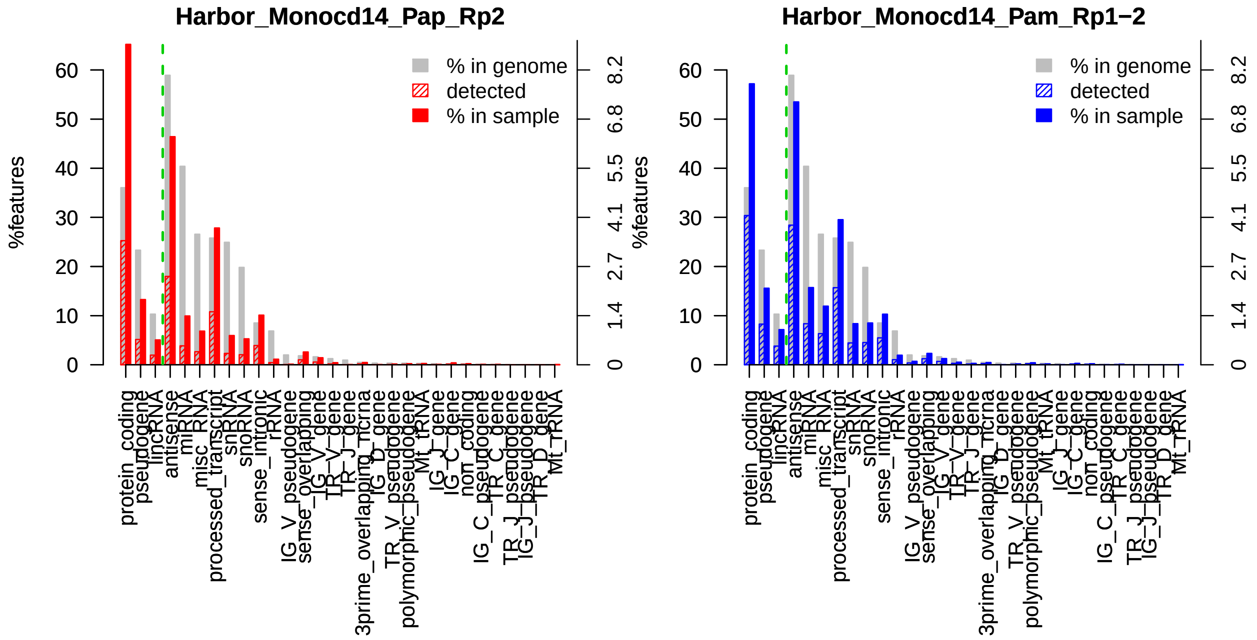
The plots in this section can be generated independently for each sample or condition of a given experiment. However, if only two samples or conditions are to be compared, the “Biotype comparison” plots (Figures S3a and S3b) can also be obtained to facilitate the interpretation and comparison of these results.



(a) Biotype detection. Pap protocol.

(b) Biotype detection. Pam protocol.

Figure S1: **Biotype distribution**. Data: *B-cells (Cd20) from ENCODE project.* Grey bars represent the percentage of each biotype in the reference genome. Stripped color bars illustrate the proportion of genes in the genome, by biotype, that are detected in the sample. Solid color bars give the percentage of each biotype within genes detected in the sample. Bars in the left hand side of the vertical green line are associated to numbers in Y left axis. Bars in the right hand side of the vertical green line are associated to numbers in Y right axis.

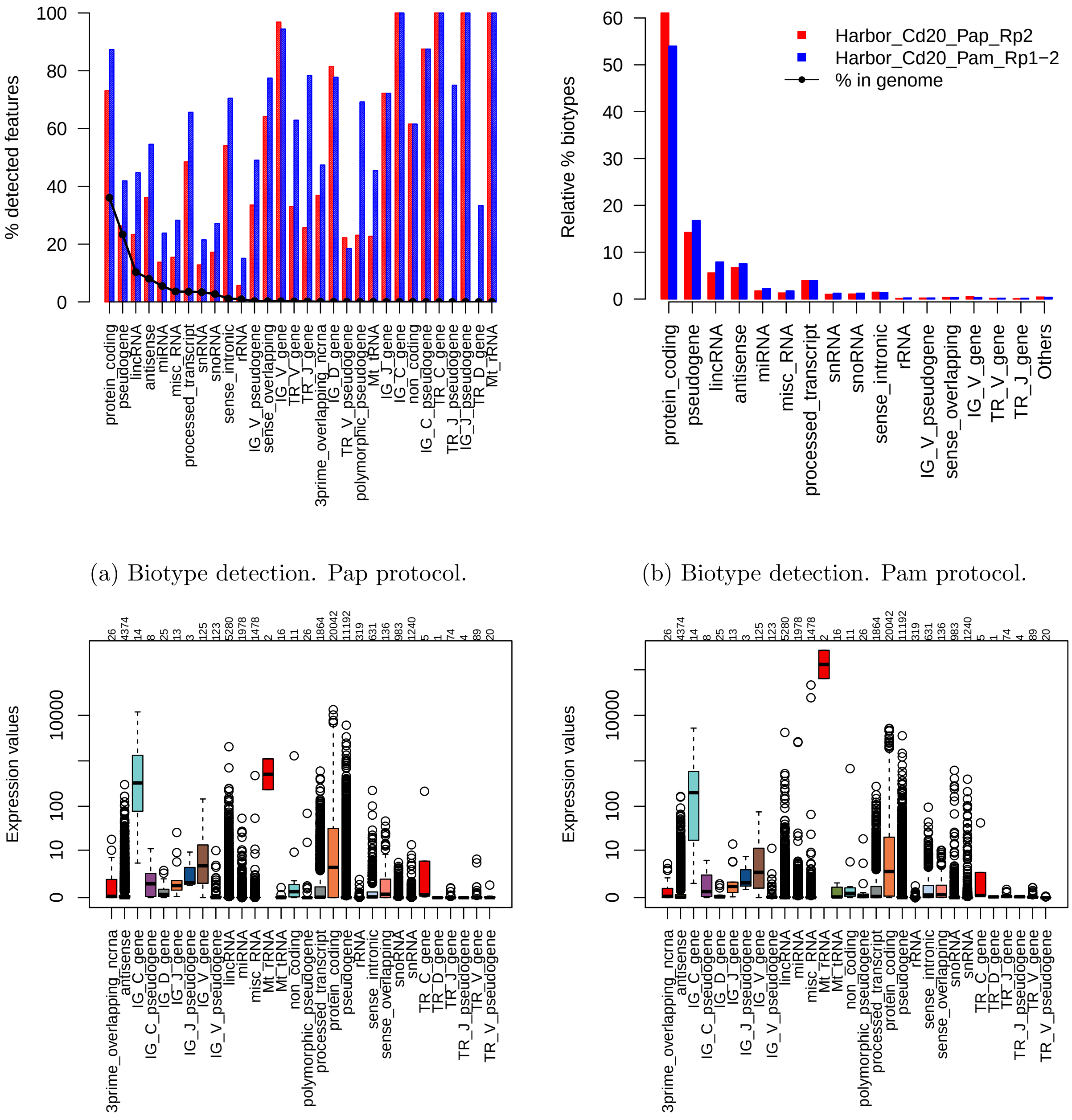


(a) Biotype detection. Pap protocol.

(b) Biotype detection. Pam protocol.

Figure S2: **Biotype distribution**. Data: *Monocytes (CD14-positive cells from human leukapheresis production) from ENCODE project.* Grey bars represent the percentage of each biotype in the reference genome. Stripped color bars illustrate the proportion of genes in the genome, by biotype, that are detected in the sample. Solid color bars give the percentage of each biotype within genes detected in the sample. Bars in the left hand side of the vertical green line are associated to numbers in Y left axis. Bars in the right hand side of the vertical green line are associated to numbers in Y right axis.

2.1.2 “Biotype comparison” and “Biotype expression” plots



(a) Biotype detection. Pap protocol.

(b) Biotype detection. Pam protocol.

(c) Biotype expression range. Pap protocol.

(d) Biotype expression range. Pam protocol.

Figure S3: **Biotype distribution**. Data: *B-cells (Cd20) from ENCODE project.* Figures S3a-S3b. Grey bars represent the percentage of each biotype in the reference genome. Stripped color bars illustrate the proportion of genes in the genome, by biotype, that are detected in the sample. Solid color bars give the percentage of each biotype within genes detected in the sample. Figures S3c-S3d. Expression values (Y axis) are given in counts per million of sequencing reads (CPM). Numbers in the upper part of the plot are the number of genes, by biotype, that are represented in each boxplot.

2.1.3 Output of explo.plot() function when asking for “Biotype comparison” plot

B-cell samples

```
 "Percentage of protein_coding biotype in each sample:"
  Harbor_Cd20_Pap_Rp2 Harbor_Cd20_Pam_Rp1-2
              61.0717               54.0017
"Confidence interval at 95% for the difference of percentages:
Harbor_Cd20_Pap_Rp2 - Harbor_Cd20_Pam_Rp1-2"
 6.4936 7.6464
"The percentage of this biotype is significantly DIFFERENT for these two samples
(p-value = 8.721e-128 )."
```

Monocyte samples

```
"Percentage of protein_coding biotype in each sample:"
  Harbor_Monocd14_Pap_Rp2 Harbor_Monocd14_Pam_Rp1-2
                  65.2175                   57.1911
"Confidence interval at 95% for the difference of percentages:
Harbor_Monocd14_Pap_Rp2 - Harbor_Monocd14_Pam_Rp1-2"
 7.4587 8.5942
 "The percentage of this biotype is significantly DIFFERENT for these two samples
(p-value = 8.094e-169 )."
```

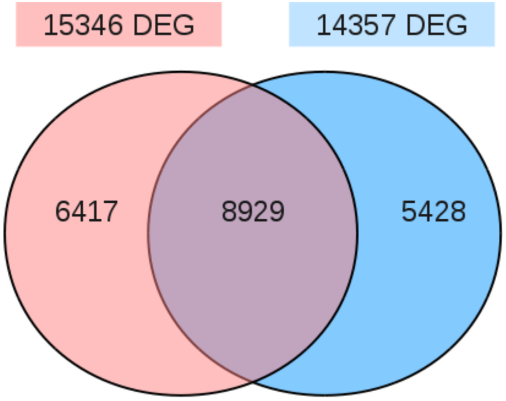


Figure S4: Number of differentially expressed genes by NOISeq-sim between B-cells and monocytes ENCODE data when using PolyA+ (pink) or Poly- (blue) extraction method.

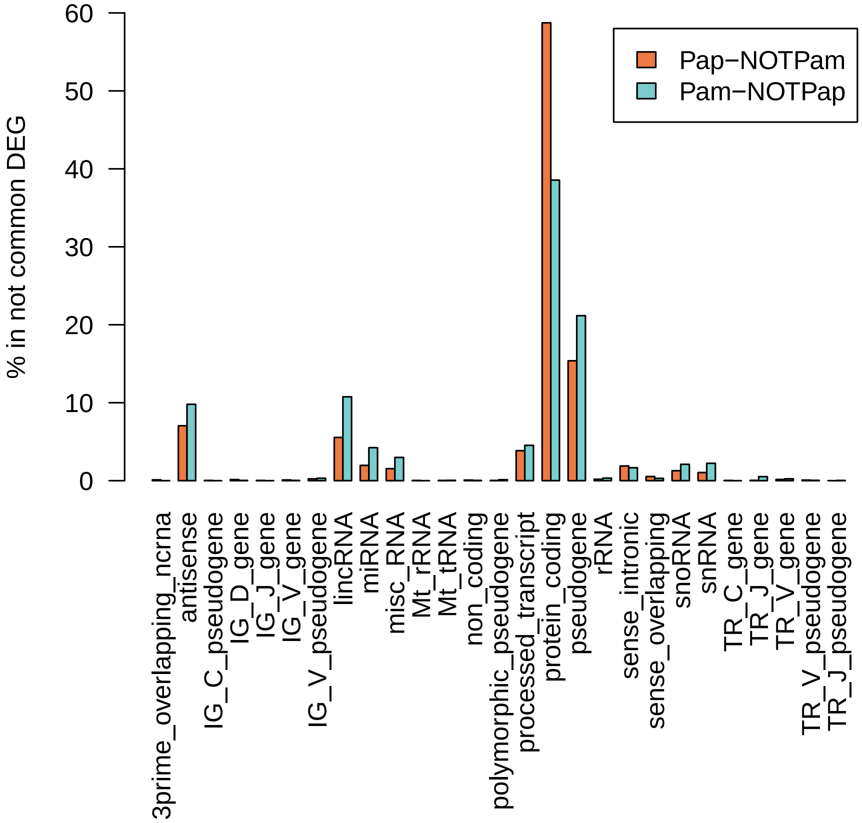


Figure S5: Percentage of each biotype within non-common differentially expressed genes between B-cells and monocytes ENCODE data when using PolyA+ or Poly- extraction method.

2.2 Sequencing depth and quantification of expression.

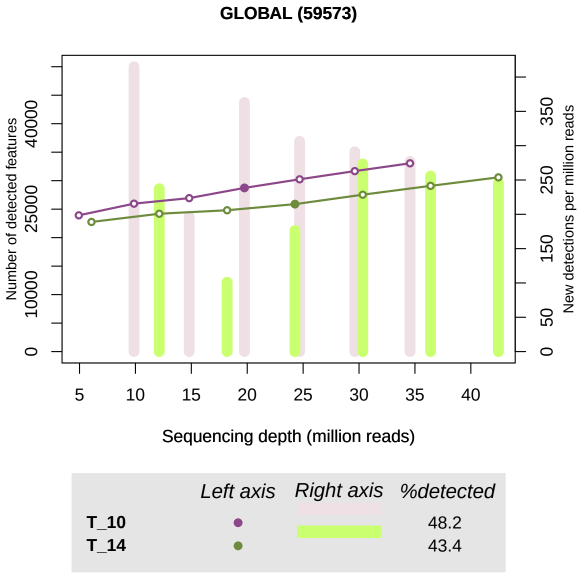


Figure S6: Saturation plot for two tumoral samples.

2.3 Sequencing biases.

2.3.1 Confidence Intervals for the median of M values

Data: *F. oxysporum* data

This is the output generated by NOISeq function to check RNA composition bias:

```
Warning: 4197 features with 0 counts in all samples are to be removed for this analysis.
Reference sample is: wt_B_30_37_1
Confidence intervals for median of M:
                 0.83%               99.17%             Diagnostic Test
wt_B_30_37_2  -0.153815574418935  -0.108393535088251  FAILED
wt_M_30_37_1  -0.370808505381945  -0.370808505381945  FAILED
wt_M_30_37_2  -0.36066807247099   -0.29410617912117   FAILED
Diagnostic test: FAILED. Normalization is required to correct this bias.
```

3 Normalization

Plot	Diagnostic test thresholds	Normalization method or *R package*
Length bias	p-value < 0.05 and $R^2 > 70\%$	*RPKM* (NOISeq), EDASeq, RNASeqBias, cqn
GC content	p-value < 0.05 and $R^2 > 70\%$	EDASeq, RNASeqBias, cqn
RNA composition	Adjusted p-value < 0.05	*TMM* (edgeR, NOISeq), Quantiles, *Upper Quartile* (EDASeq, NOISeq), DESeq, DESeq2

4 Filtering out low count features

It has been often argued that, in RNA-seq, expression estimation for low count genes is less reliable because read counts could have been assigned by chance [1, 2]. Thus, excluding features with low counts may improve the results of statistical analyses because the level of noise is reduced. However, the best procedure to filter these low count features has not yet been decided.

These filtering procedures have not been implemented in statistical packages for RNA-seq data but it is a common practice simply removing genes with total counts for all the samples lower than a certain cutoff, e.g. 10 counts [3–5]. This approach does not take into account the sequencing depth of the experiment to decide the cutoff, so genes with a relatively high expression in one of the conditions could be ignored. A better method is the procedure described in the edgeR R package User's Guide in the Bioconductor repository. The authors proposal consists of keeping genes with counts per million reads (CPM) above a given threshold in at least as many samples as the number of samples per condition. By setting the cutoff for the counts per million instead of the raw counts, it can be assured that no genes with a high relative expression are eliminated. In the NOISeq package, we implemented three different filtering procedures: CPM method, Wilcoxon test and Proportion test, that are described in detail in the next sections.

4.1 CPM method

Let x_g^s be the number of raw counts of gene g in sample s. As in edgeR proposal, counts for each sample are transformed to counts per million reads (CPM): $CPM_g^s = 10^6 \times \frac{x_g^s}{\sum\limits_g x_g^s}$. A value for CPM under which a feature is considered to have low counts must be previously set (*cpm*). By default, CPM method takes a cutoff of $cpm = 1$. If there are S samples in a given experimental condition, the cutoff for that condition would be $cpm \times S$. A gene g is filtered out if the sum of CPM values across all samples in the same condition is below the condition cutoff ($\sum\limits_s CPM_g^s < cpm \times S$) for all the experimental conditions.

It is also possible to remove genes that present inconsistent expression values in any of the experimental conditions with the CPM method. A cutoff for the coefficient of variation per condition *cv* has to be set a priori. Then, a gene g will be filtered out either if it has a total CPM value per condition of less than $cpm \times S$ or a coefficient of variation per condition higher than the *cv* cutoff for all the conditions.

4.2 Wilcoxon test

Although the CPM method takes the experimental design and the variability per condition into account, it has the drawback of having to decide the cutoffs to use for both the CPM and the coefficient of variation. Hence, we propose the Wilcoxon test to identify those genes with a CPM value median per condition that is significantly higher than 0. Thus, the hypothesis to test for each gene and condition is $H_0 : m = 0$ versus $H_1 : m > 0$, where m is the median of CPM values per condition. To be more conservative, no multiple testing correction was applied in order to retain as many genes as possible. Genes with a p-value higher than 0.05 in all the conditions are filtered out.

By using the Wilcoxon method, genes with inconsistent values across replicates within the same condition or with a low median expression value tend to be removed. However, this non-parametric procedure is only recommended when the number of replicates per condition is at least 5.

4.3 Proportion test

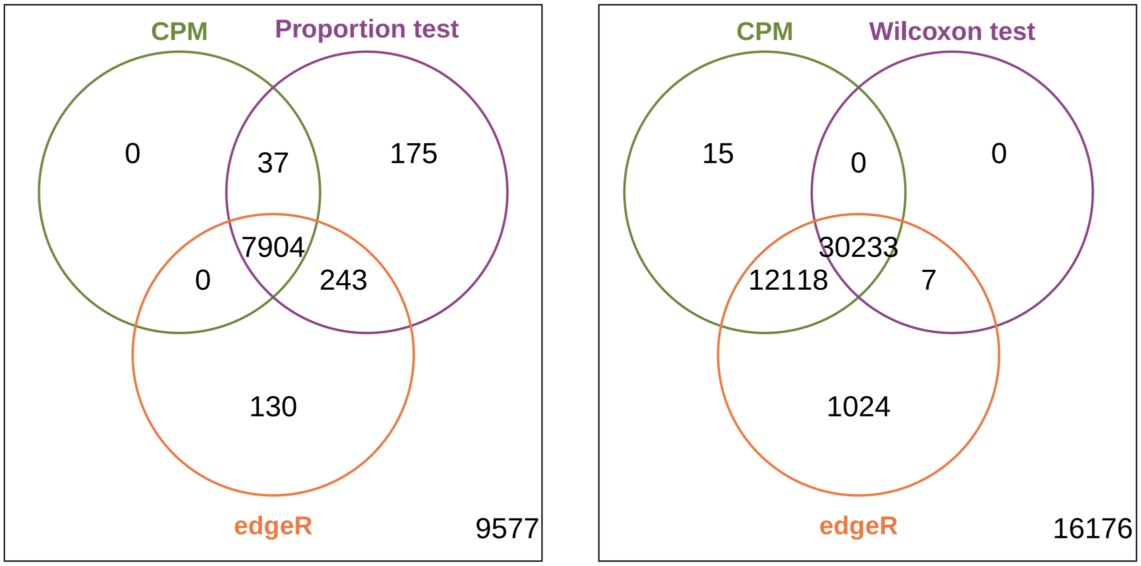
The proportion test aims to be the alternative to the Wilcoxon test when few replicates per condition are available. This method requires a cutoff to be set for CPM (cpm), but not for the coefficient of variation. It is based on the idea that read counts for a given gene follow a binomial distribution where the number of trials n is the sequencing depth, and the probability p, is the probability of expression for that gene under a given experimental condition, which is unknown. Thus, in this case, $H_0 : p = p_0$ is tested versus $H_1 : p > p_0$. Since it is not possible to use $p_0 = 0$ in a binomial proportion test, we define $p_0 = cpm/10^6$. If several replicates are available for an experimental condition, we sum across replicates ($x_g = \sum\limits_s x_g^s$) and use this single value as the observed binomial variable. Then, $n = \sum\limits_g x_g$. Again, to be conservative, the raw p-values are used and genes with a p-value higher than 0.05 in all conditions are filtered out.

4.4 Comparing filtering methods

We applied the three NOISeq filtering procedures and edgeR proposal to *F. oxysporum* (with 2 replicates per condition) and Prostate Cancer data (with 11 and 12 replicates per condition) to illustrate the similarities and differences of the methods. We set a cutoff of $cpm = 1$ for CPM method, Proportion test, and edgeR approach. Because of the number of replicates, the Proportion test was only applied to *F. oxysporum* and the Wilcoxon test was only applied to Prostate Cancer data. We considered a coefficient of variation of 500 for the CPM method to cancel this filter and make this method more comparable to edgeR approach. According to the number of replicates per condition in each data set, genes with CPM higher than 1 in at least 2 or 10 samples for each data set respectively were retained in the edgeR approach.

Both data sets originally contained 18066 (*F. oxysporum*) and 59573 (Prostate Cancer) genes. Out of these, 9577 and 16176 respectively, were not filtered out by any of the methods (Figure S7). Most of the filtered genes (7904 and 30233) were removed by all the methods which indicates that, in general, there were very few differences among them. The greatest difference was found for the Wilcoxon test (Prostate Cancer), since there were more than 12000 that were removed by CPM and edgeR but not by Wilcoxon.

We studied the characteristics of the removed genes that were not in common for the compared filtering methods by plotting the difference between the mean CPM per condition against the maximum variability between replicates (Figures S8 and S9). In general, genes filtered only by edgeR tended to show higher differences in expression between conditions which is obviously not



(a) *F. oxysporum* data

(b) Prostate Cancer data

Figure S7: Number of genes filtered out by each method

good because genes with potentially significant changes in expression between conditions could be removed from the analysis. Although these genes generally present a high variability among replicates and will probably not be declared as differentially expressed by statistical methods, it may be preferable to leave the decision about these cases to the statistical method instead of filtering them out of the ulterior analysis.

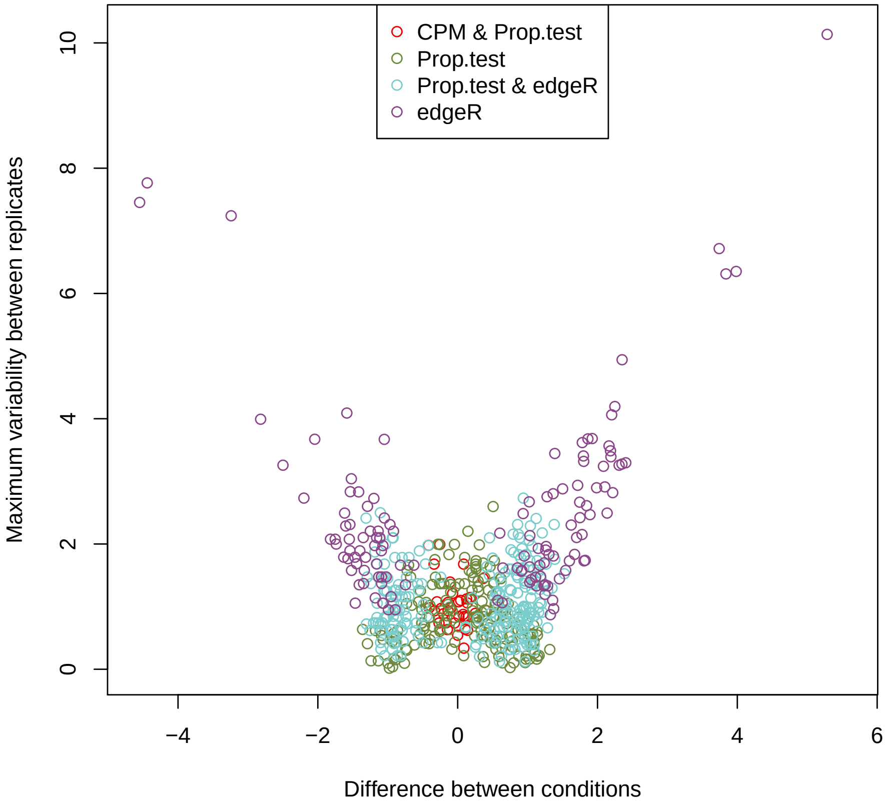
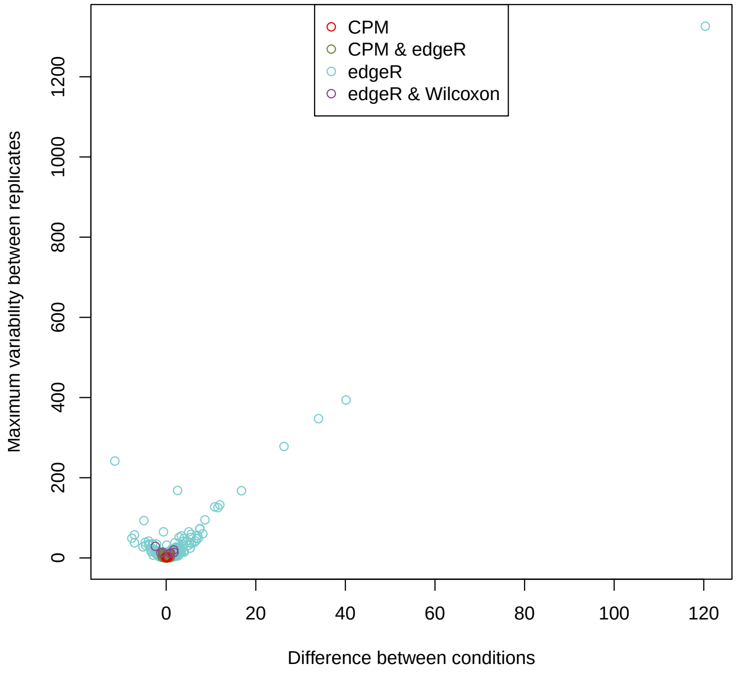
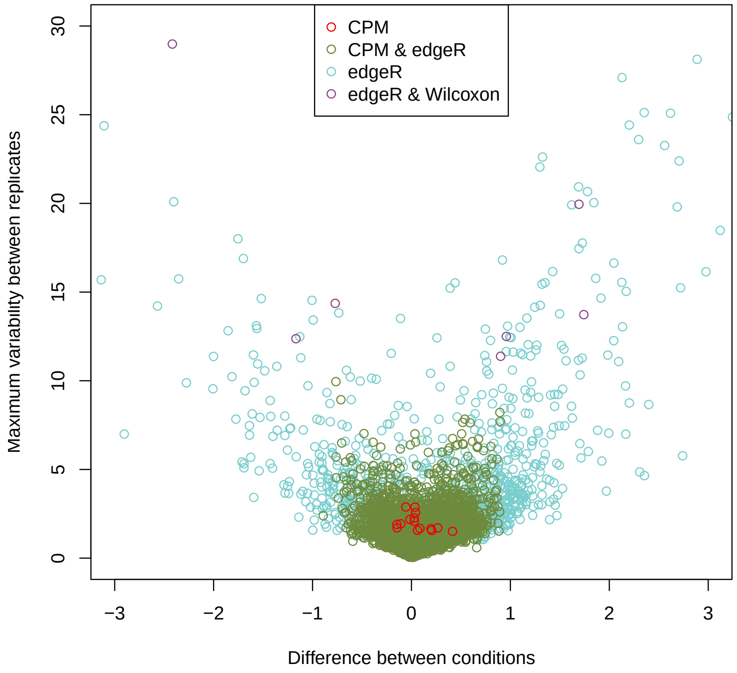


Figure S8: Difference between expression mean per condition versus maximum difference between replicates for genes not removed by all methods from *F. oxysporum* data.



(a) All data



(b) Zoomed data

Figure S9: Difference between expression mean per condition versus maximum difference between replicates for genes not removed by all methods from Prostate Cancer data.

5 Differential expression

5.1 Results on experimental data sets

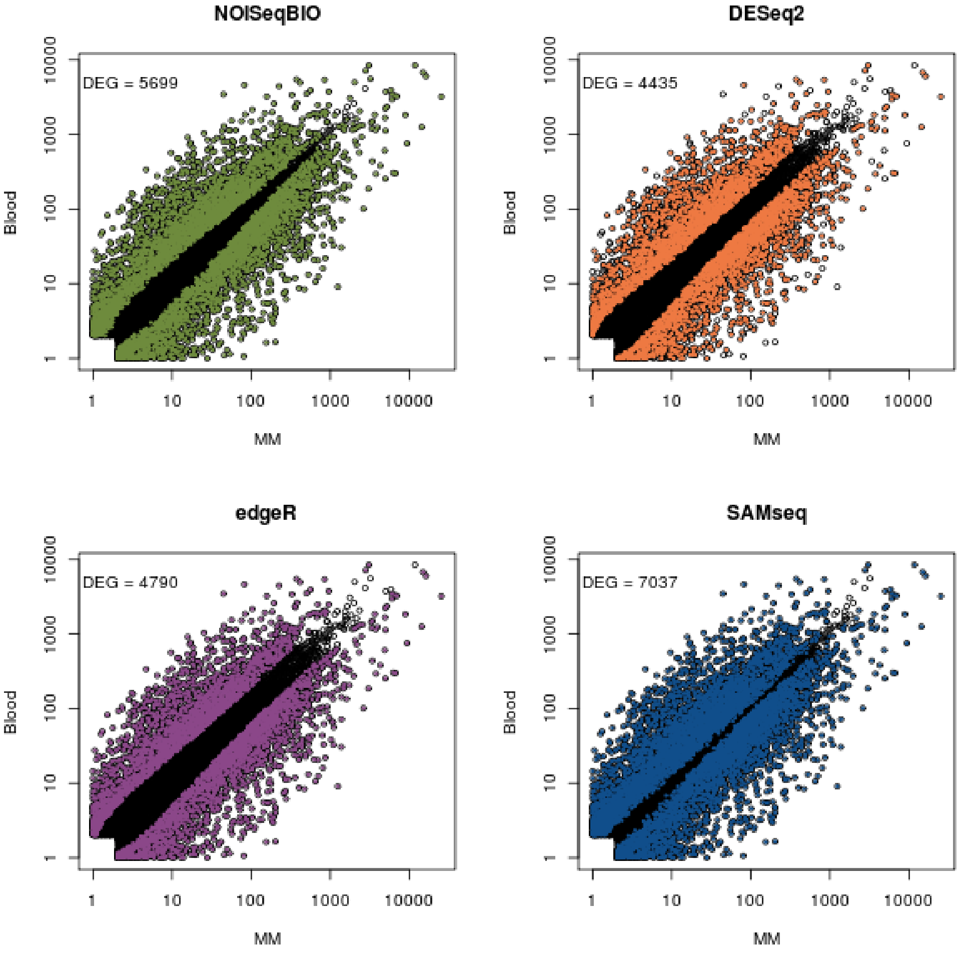


Figure S10: Differential expression results from compared methods on *F. oxysporum* data. The DEGs declared by each method are shown in color.

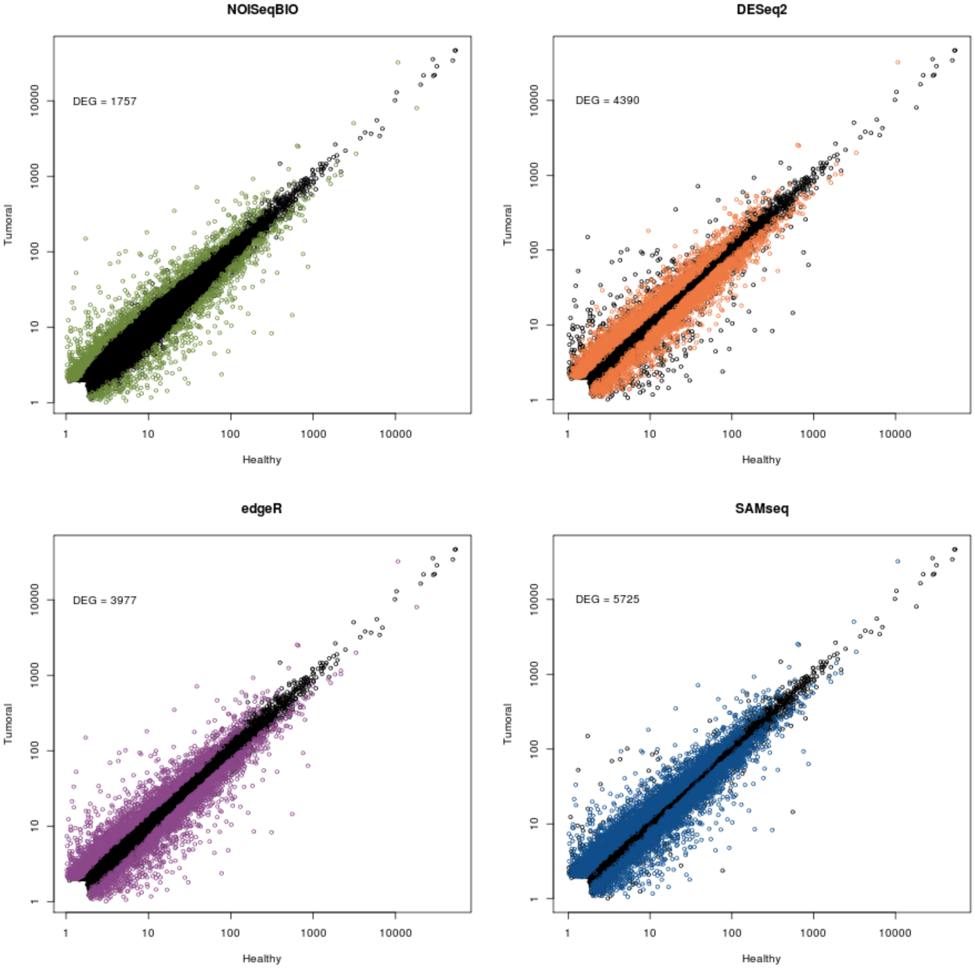
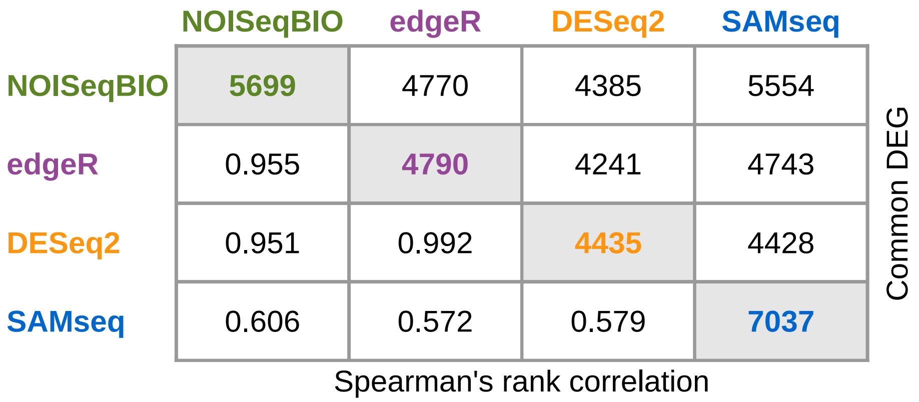


Figure S11: Differential expression results from compared methods on human prostate data. The DEGs declared by each method are shown in color.

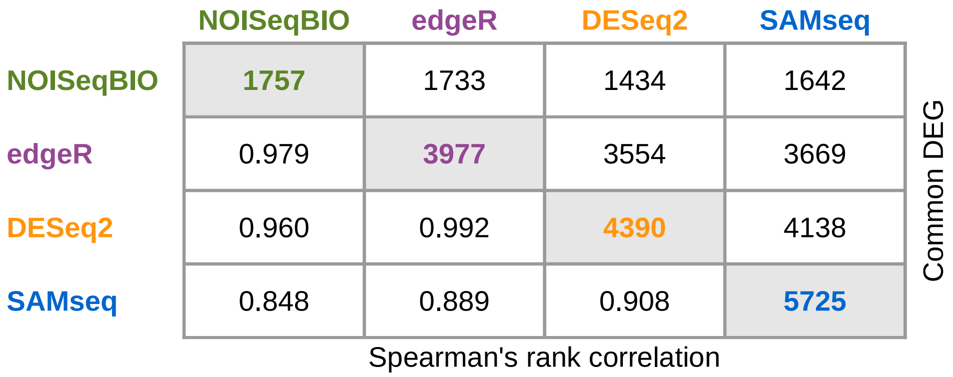


	NOISeqBIO	edgeR	DESeq2	SAMseq
NOISeqBIO	5699	4770	4385	5554
edgeR	0.955	4790	4241	4743
DESeq2	0.951	0.992	4435	4428
SAMseq	0.606	0.572	0.579	7037

Common DEG

Spearman's rank correlation

Figure S12: Differential expression results from *F. oxysporum* data. The diagonal contains the number of DEGs for each method. Above the diagonal, the number of DEGs in common for each pair of methods is shown. Below the diagonal, the Spearman's rank correlation coefficient between FDR or 1-probability for each pair of methods is shown.



	NOISeqBIO	edgeR	DESeq2	SAMseq
NOISeqBIO	1757	1733	1434	1642
edgeR	0.979	3977	3554	3669
DESeq2	0.960	0.992	4390	4138
SAMseq	0.848	0.889	0.908	5725

Common DEG

Spearman's rank correlation

Figure S13: Differential expression results from human prostate cancer data. The diagonal contains the number of DEGs for each method. Above the diagonal, the number of DEGs in common for each pair of methods is shown. Below the diagonal, the Spearman's rank correlation coefficient between FDR or 1-probability for each pair of methods is shown.

5.2 Results on simulations

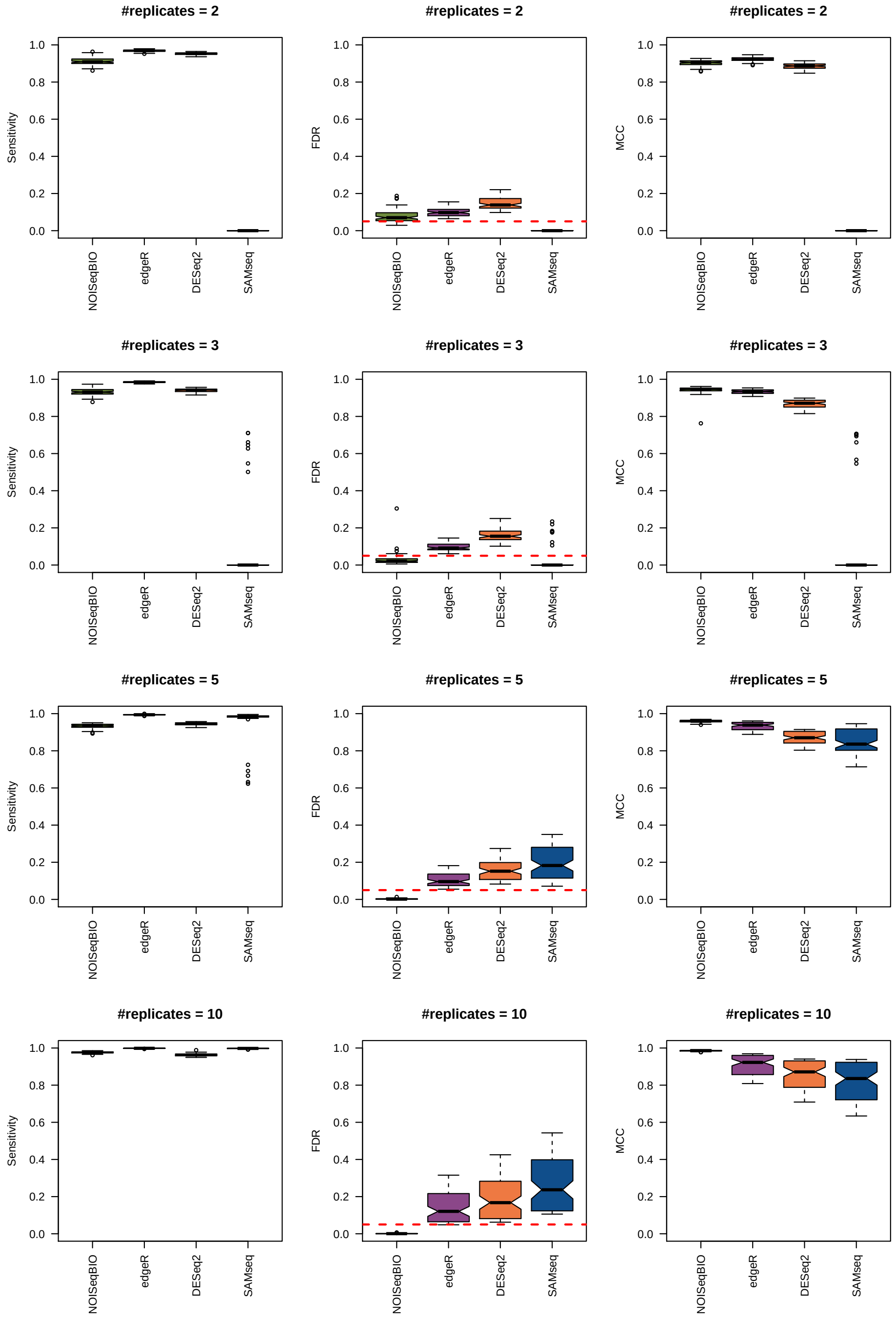


Figure S14: HIGH biological variability scenario (320 simulations). Sensitivity (left), FDR (middle), and MCC (right) of differential expression methods according to the number of replicates for an adjusted p-value cutoff of 0.05 (equivalent to a probability of 0.95 for NOISeqBIO). This cutoff corresponds to the red horizontal line in FDR plots. Results for all levels of technical noise, DEG proportions, and number of genes were aggregated.

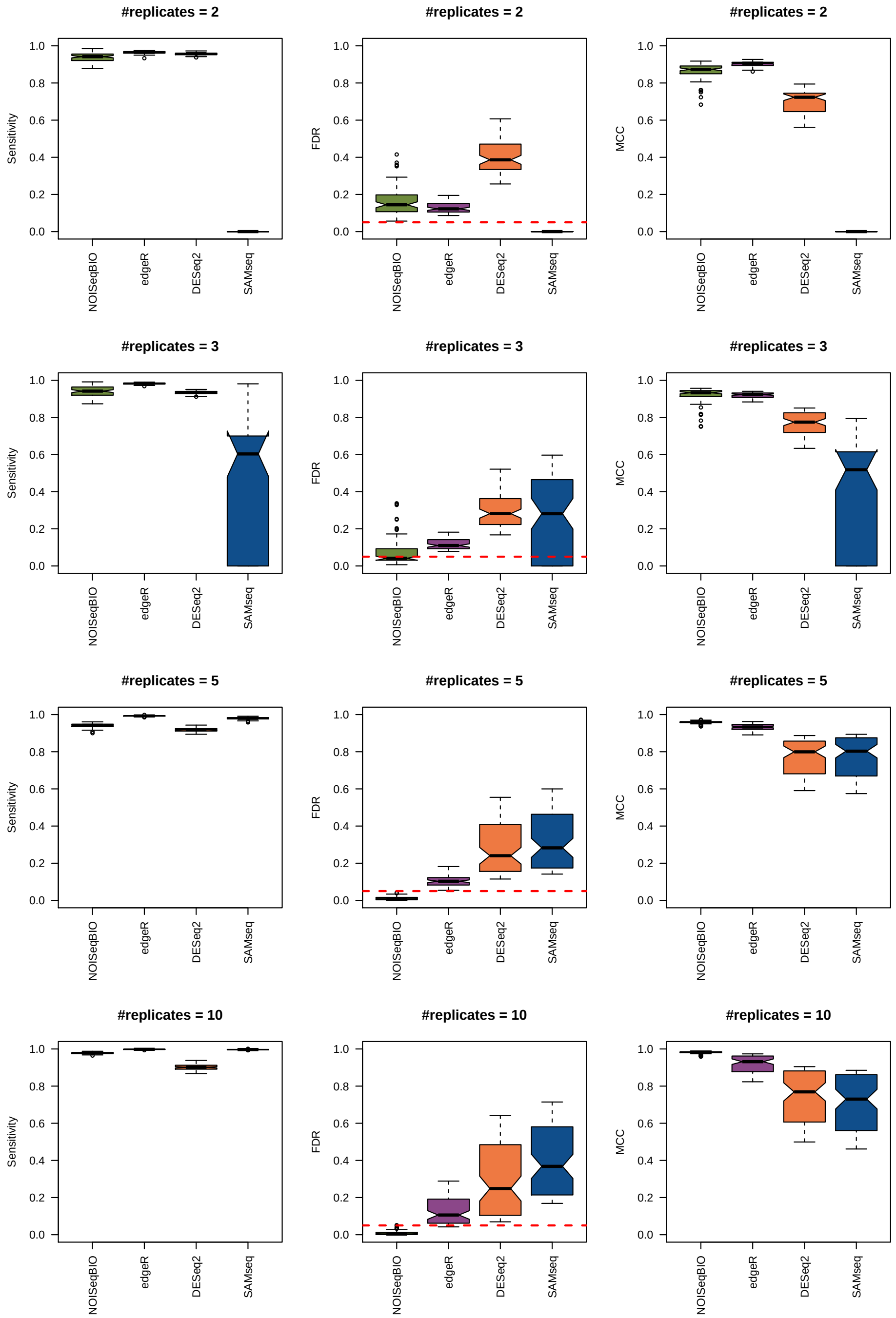


Figure S15: LOW biological variability scenario (320 simulations). Sensitivity (left), FDR (middle), and MCC (right) of differential expression methods according to the number of replicates for an adjusted p-value cutoff of 0.05 (equivalent to a probability of 0.95 for NOISeqBIO). This cutoff corresponds to the red horizontal line in FDR plots. Results for all levels of technical noise, DEG proportions, and number of genes were aggregated.

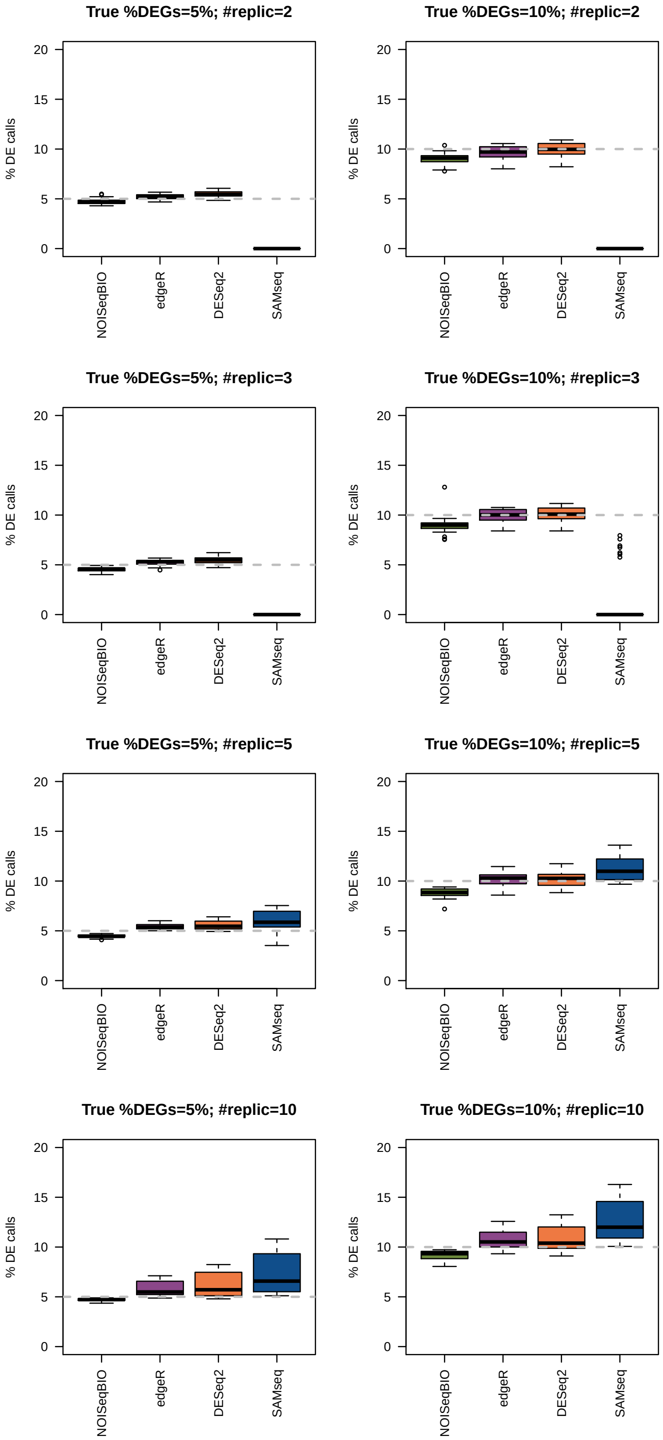


Figure S16: Percentage of differentially expressed genes called by each method with regard to the total number of simulated genes in the HIGH variability scenario,per number of replicates (in rows) and true percentage of differentially expressed genes simulated (in columns). Grey horizontal line indicates the true percentage of differentially expressed genes simulated.

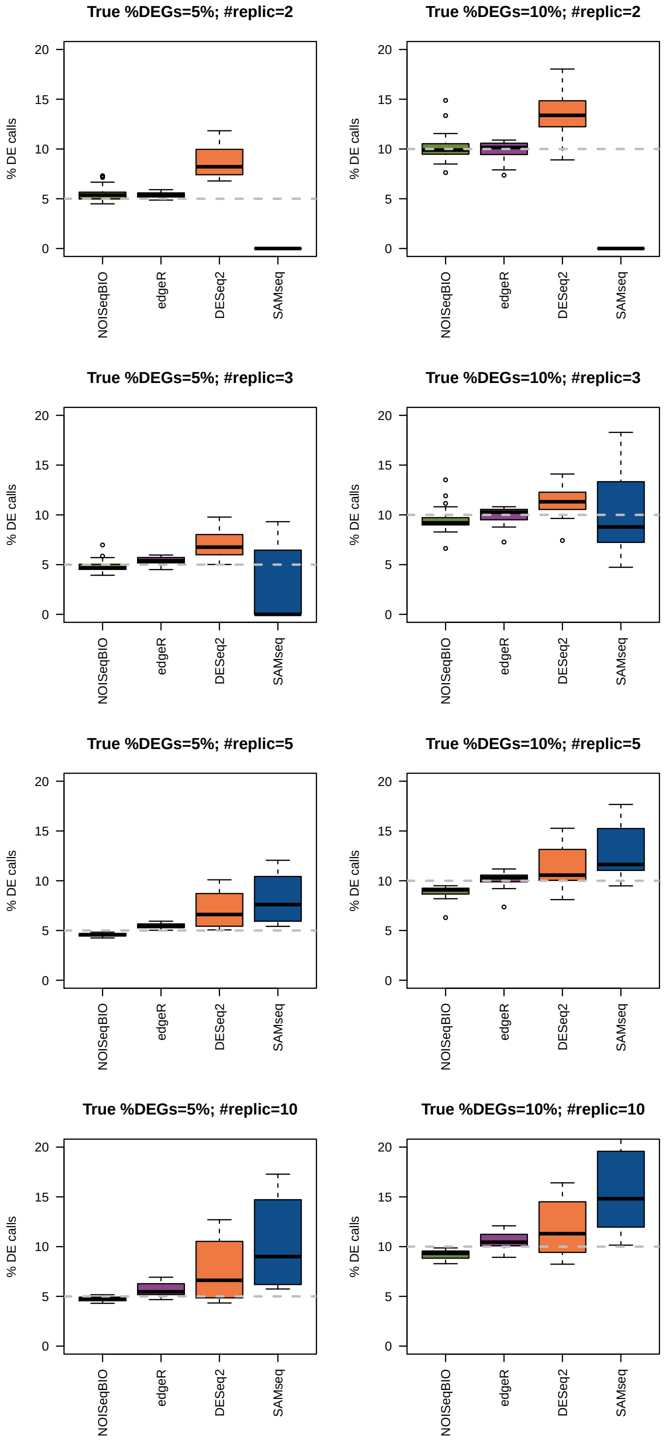


Figure S17: Percentage of differentially expressed genes called by each method with regard to the total number of simulated genes in the LOW variability scenario,per number of replicates (in rows) and true percentage of differentially expressed genes simulated (in columns). Grey horizontal line indicates the true percentage of differentially expressed genes simulated.

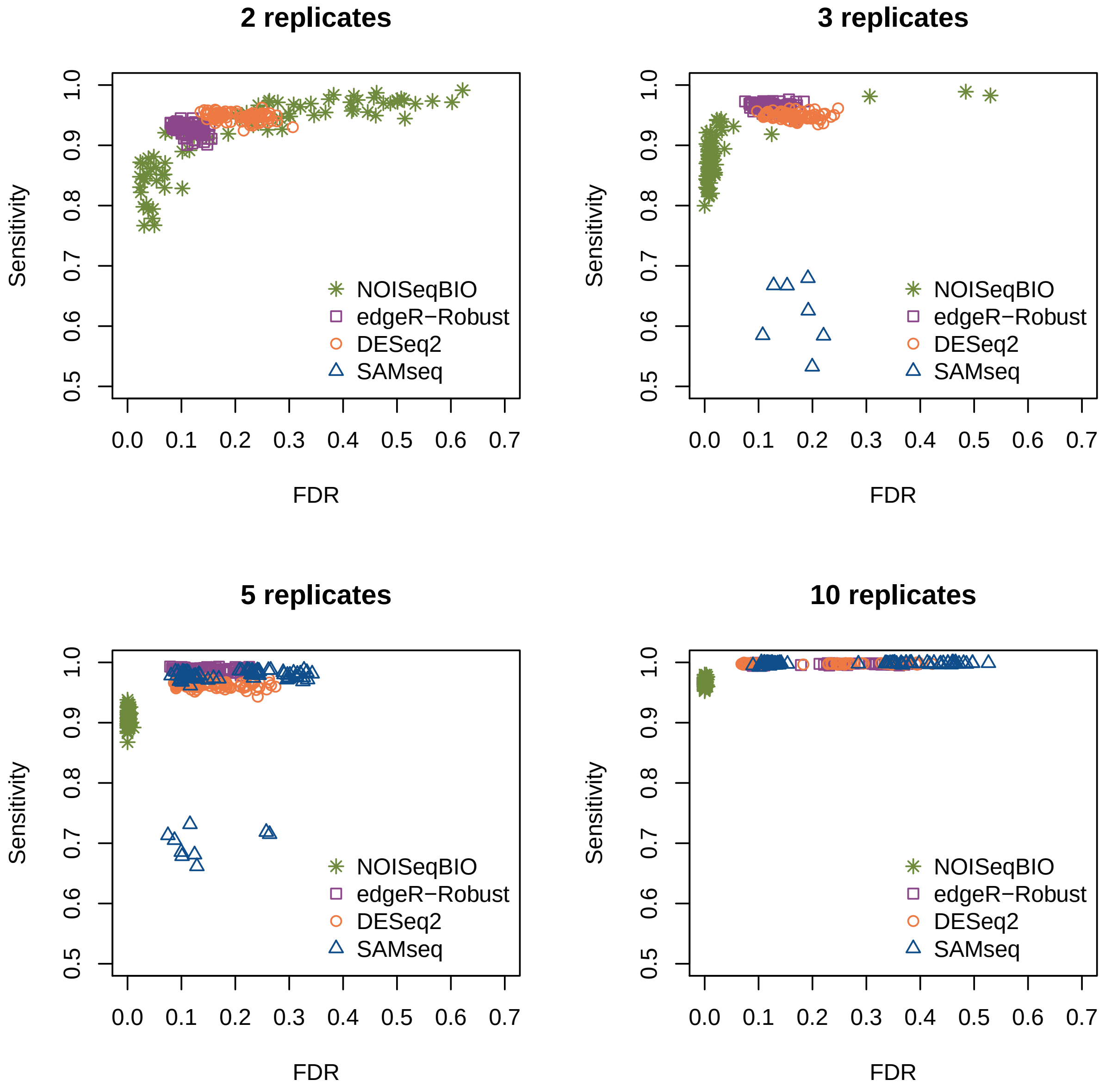


Figure S18: Trade-off between FDR and sensitivity for each DE method at a significance level of 5% in the HIGH variability scenario (320 simulations) with 10% of genes presenting an outlier sample (original simulated value times a number between 5 and 10), as done in [4]. Please note that for SAMseq there are results below the axis limits that are not displayed for the sake of clarity.

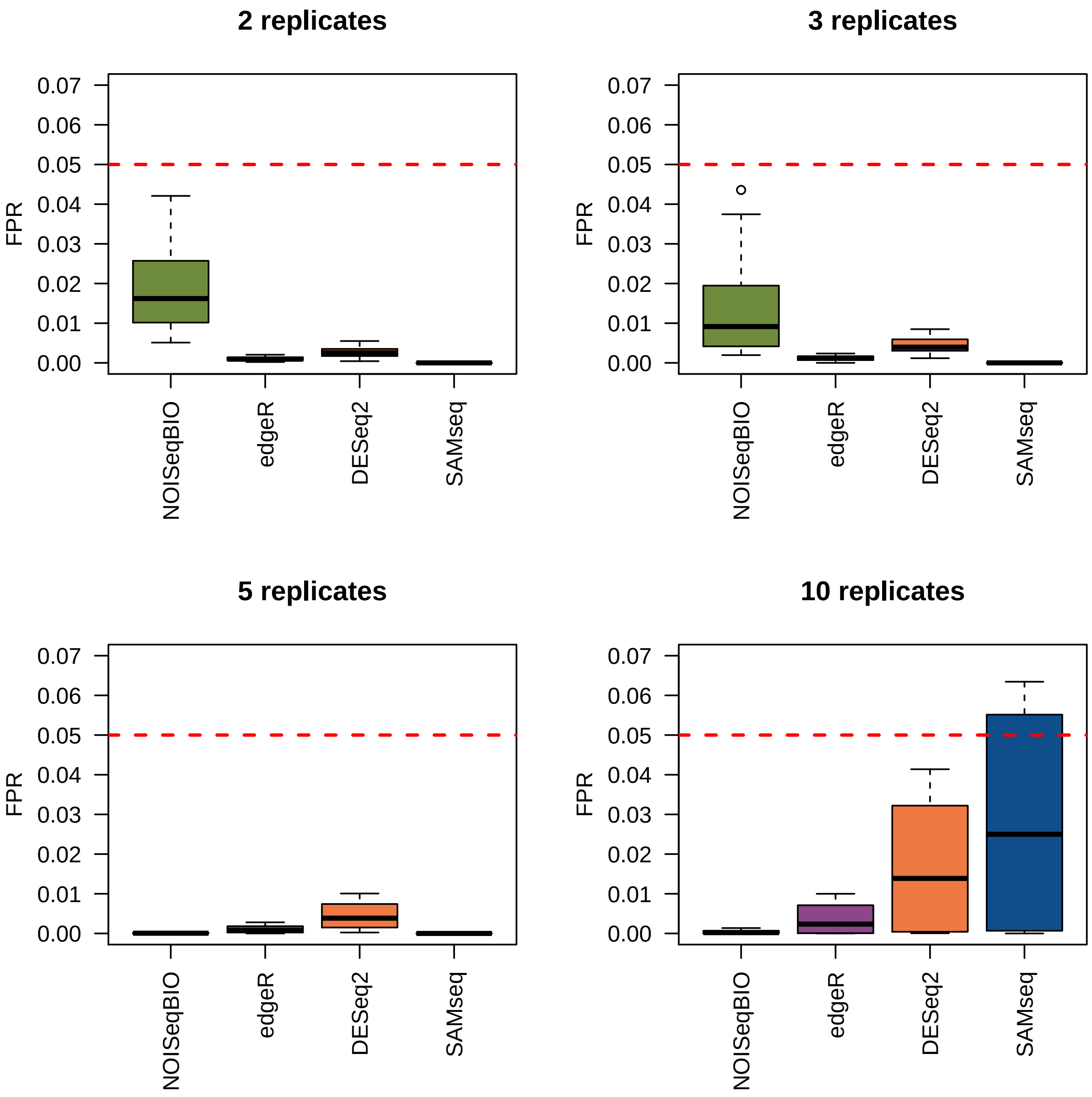
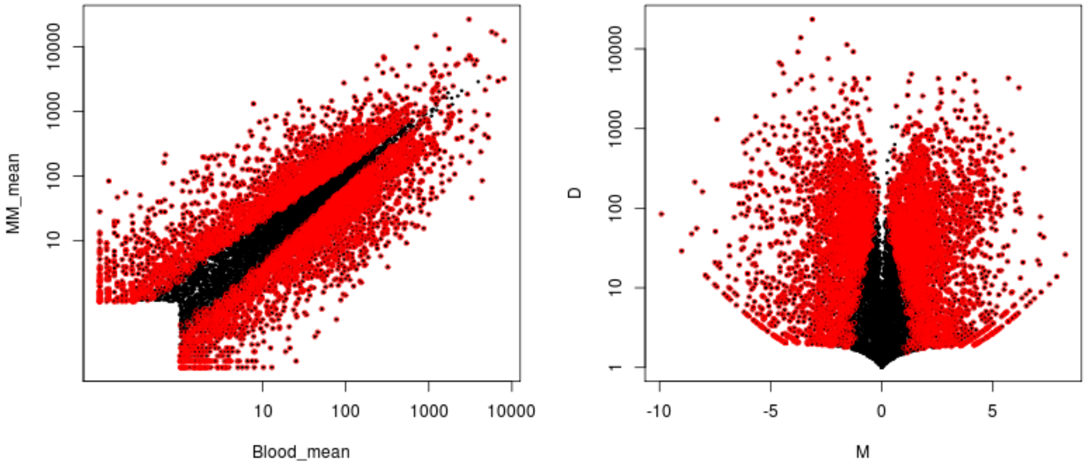


Figure S19: False Positive Rate (FPR) per number of replicates and for each method applied on simulations with 0 DEGs at 5% significance level. 40 simulations were generated for each number of replicates. Red horizontal line show the significance level (5%).

5.3 Visualization of DE results

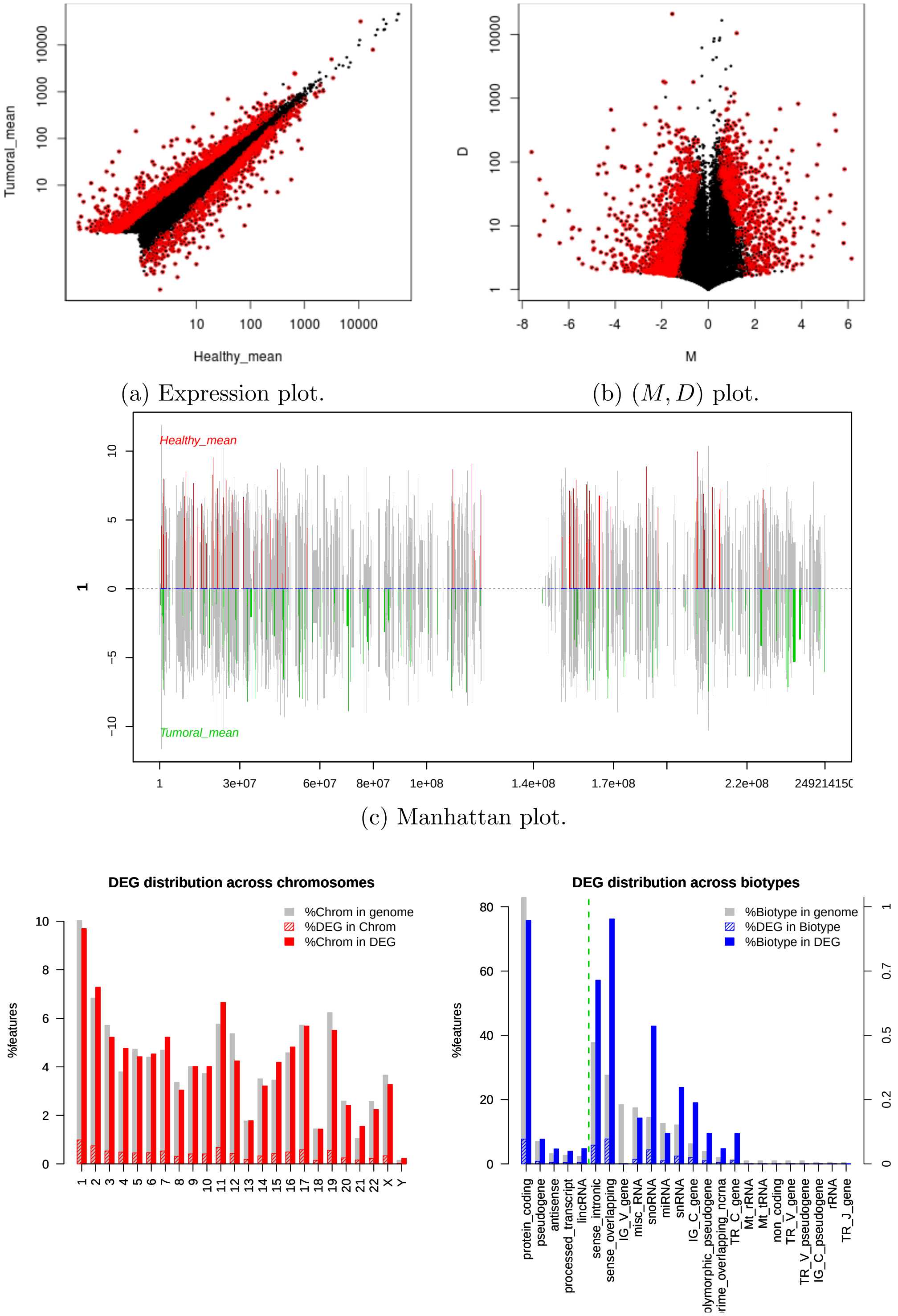
The NOISeq package also includes a wide variety of plots to visualize DE results that can provide additional insights for the biological interpretation of the significant gene calls made (Figures S20, S10 and S11). Please, refer to the package User's Guide (http://www.bioconductor.org/packages/release/bioc/html/NOISeq.html) for a detailed explanation of all of these plots. "Expression" and "(M, D)" plots (Figures S21a and S10a, and S21b and S10b) allow for the detection of any bias in DEGs, for example, only declaring high or low count genes or genes with a high fold-change, as DEGs. No biases were observed in our DE results. "Manhattan" plots (Figures S10c and S11) display the expression level across all the chromosomal positions and indicate up and down regulated DEGs. In the Prostate Cancer data set (Figure S10c), we observe a region of extensive gene expression up-regulation on the left arm of Chromosome 1. On the other hand, the DEG chromosome-break down for the *F. oxysporum* data set (Figure S11) revealed reduced gene expression on chromosomes 3,6,14, and 15. Interestingly, these chromosomes have recently been introduced into the fungal genome suggesting that the observed gene silencing pattern may reflect an ongoing adaptive process of these chromatin parts. Finally, the package also provides the biotype or chromosome break-down for DE genes. For example, for the cancer data (Figure S10d), the biotype plot reveals that although these DEGs are protein-coding genes, there is a relatively higher proportion of DE non-coding genes than expected from the general genome composition (solid blue bars are higher than gray bars for these biotypes), suggesting that non-coding RNA regulatory mechanisms might be specifically activated in this experimental setting.



(a) Expression plot.

(b) (M, D) plot.

Figure S20: **Differential Expression plots**. Data: *F. oxysporum*.



(a) Expression plot.

(b) (M, D) plot.

(c) Manhattan plot.

(d) Distribution of DEG across chromosomes and biotypes.

Figure S21: **Differential Expression plots**. Data: Prostate cancer.

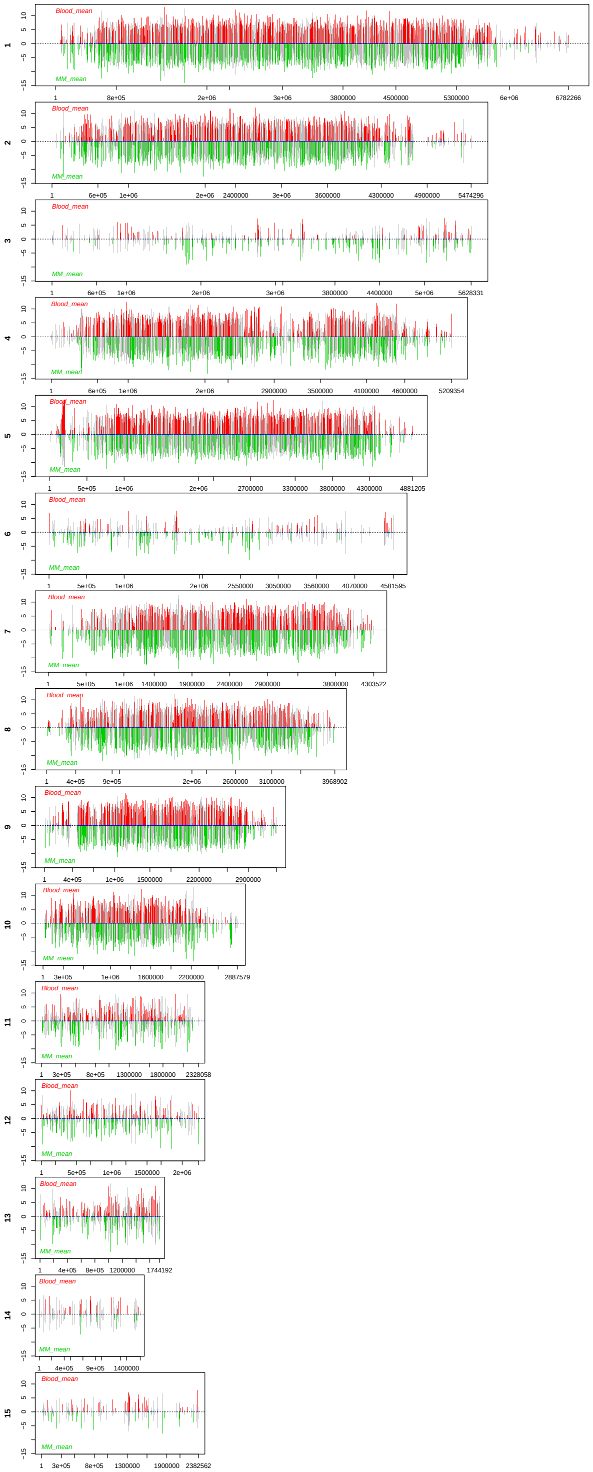


Figure S22: **Manhattan plot**. Data: *F. oxysporum*.

6 Simulation algorithm

It has been reported [5–7] that the number of reads mapping to a given gene resembles an over-dispersed Poisson distribution when considering biological replicates and that one way of modeling this over-dispersion is by taking the negative binomial distribution. Thus, our simulation algorithm is based on randomly generating the counts from a negative binomial distribution as done previously in other studies [4, 8]. Figure S23 shows the outline of the algorithm.

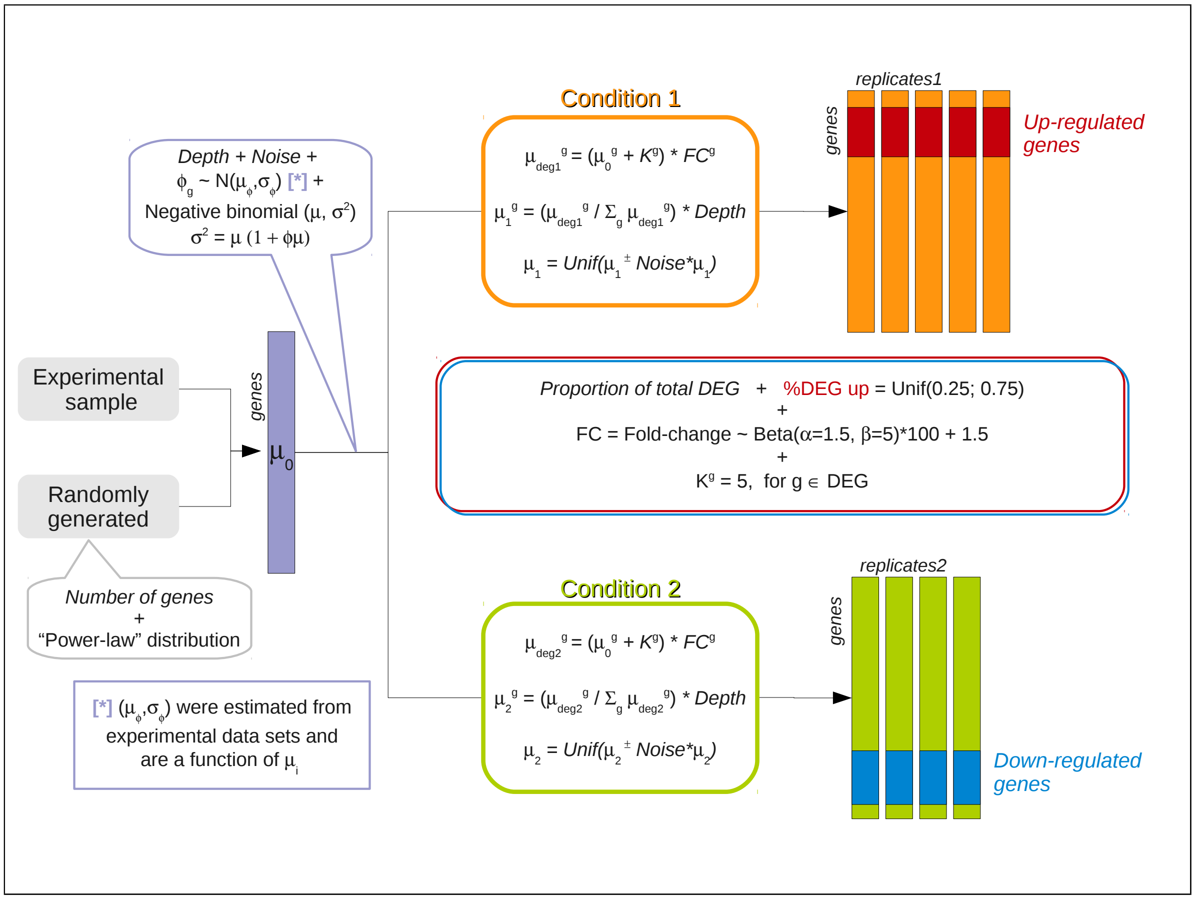

Condition 1
replicates1
genes
Up-regulated genes
$\mu_{deg1}{}^{g} = (\mu_0{}^{g} + K^g) * FC^g$
$\mu_1{}^{g} = (\mu_{deg1}{}^{g} / \Sigma_g \, \mu_{deg1}{}^{g}) * Depth$
$\mu_1 = Unif(\mu_1 \pm Noise*\mu_1)$
Depth + Noise +
$\phi_g \sim N(\mu_\phi,\sigma_\phi)$ [*] +
Negative binomial (μ, σ^2)
$\sigma^2 = \mu\,(1 + \phi\mu)$
Experimental sample
Randomly generated
genes
μ_0
Number of genes
+
“Power-law” distribution
Proportion of total DEG + %DEG up = Unif(0.25; 0.75)
+
FC = Fold-change ~ Beta(α=1.5, β=5)*100 + 1.5
+
$K^g = 5$, for $g \in$ DEG
Condition 2
replicates2
genes
Down-regulated genes
$\mu_{deg2}{}^{g} = (\mu_0{}^{g} + K^g) * FC^g$
$\mu_2{}^{g} = (\mu_{deg2}{}^{g} / \Sigma_g \, \mu_{deg2}{}^{g}) * Depth$
$\mu_2 = Unif(\mu_2 \pm Noise*\mu_2)$
[*] (μ_ϕ,σ_ϕ) were estimated from experimental data sets and are a function of μ_i

Figure S23: Outline of the simulation algorithm

These are the main steps of the simulation algorithm:

1. An initial number of counts per gene (μ_0) is used to simulate the replicates for each condition. This μ_0 determines the proportion of sequencing reads initially assigned to each gene. It can be either provided by the user or randomly generated from a power-law distribution: $f(x) \propto x^{-\lambda}$, where $0 \leq x \leq depth/1000$ and $\lambda = 0.5$. Thus, if $ngenes$ is the number of genes in the simulated data set, $ngenes$ values are randomly generated from this distribution to be used as the initial counts μ_0 when no experimental samples are provided by the user.

2. The proportion of differentially expressed genes ($propdeg$) is chosen by the user and is used to obtain the number of DEGs. The proportion of DEGs that will be up-regulated in condition 1 is generated from the uniform distribution $\mathcal{U}(0.25, 0.75)$, and the rest of the DEGs are down-regulated in this condition. Genes that are up and down regulated are randomly taken from the total set of genes.

3. The number of replicates per condition is given by the parameter *nrepl*. Each biological replicate for a given gene and condition is simulated from a negative binomial distribution with mean μ and variance σ^2. To describe the relationship between the mean and the variance, the parametrization used in [6] was applied: $\sigma^2 = \mu(1 + \phi\mu)$. This is how μ and σ^2 are estimated from the initial counts μ_0:

   - For each condition i ($i = 1, 2$), the mean expression is defined as $\mu_i^g = (\mu_0^g + K^g) \times FC^g$, for $g \in DEG$, and $\mu_i^g = \mu_0^g$, for $g \notin DEG$. The fold-change FC^g is randomly generated from a Beta distribution: $\dfrac{FC^g - 1.5}{100} \sim Beta(\alpha, \beta)$, where $\alpha = 1.5$. By default, $\beta = 6$, but it can be modified by the user (see Figure S24). $K^g = 5, \forall g \in DEG$. The mean μ_i thereby obtained for each condition is adjusted so its sum is equal to the given total number of counts *depth*. Finally, in order to allow a certain level of noise in the data (*noise*), the final μ_i^g is the maximum between 0.1 and a random value from the uniform distribution $\mathcal{U}(\mu_i - noise \times \mu_i, \mu_i + noise \times \mu_i)$. The reason for taking the maximum is to give any gene with no initial counts some chance to appear.
   - To compute the variance σ^2, we first need to estimate the value of the dispersion parameter ϕ. We evaluated several experimental data sets with different number of replicates and biological variability to obtain realistic scenarios of either high or low biological variability. We followed the estimation procedure described in [8]. For each data set, only the samples with a total number of counts higher than 10^6 and the genes with a mean expression higher than 1 were chosen. Once this filter was applied, the remaining samples were adjusted so all of them had the same number of counts (depth). With these normalized data, the mean expression of each gene was computed, which is the maximum likelihood estimator (MLE) of μ^g. The MLE of ϕ^g was obtained by maximizing the log-likelihood function. This was done for each experimental data set and all the pairs (μ^g, ϕ^g) from every data set were pooled. All μ values were divided into bins containing approximately 1000 values each. Figure S25 shows the dependence of ϕ^g on μ^g for the scenario of high biological variability. The higher the value of μ, the lower the median and variability of ϕ. The mid-point of the bin was computed for each bin of values, as well as the median and the median absolute deviation (mad) of ϕ values within the bin. Thus, for each condition i, ϕ^g is randomly taken from a normal distribution $N(\mu_\phi^g, \sigma_\phi^g)$, where $\mu_\phi^g = \mu_\phi(\mu_i^g)$ and $\sigma_\phi^g = \sigma_\phi(\mu_i^g)$ are obtained by linear interpolation from μ mid-points and ϕ medians.

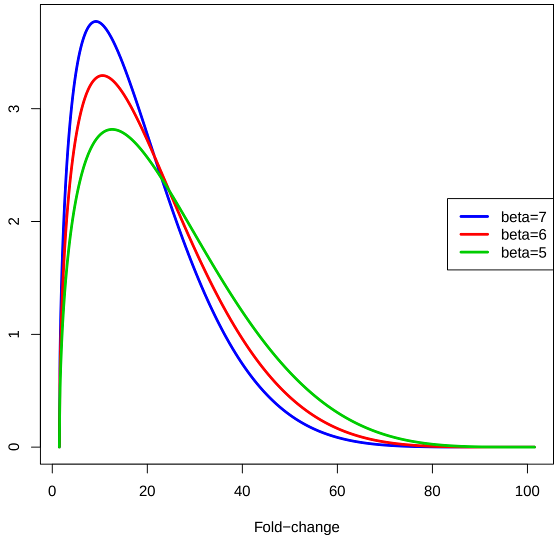


Figure S24: Fold-change is generated randomly from a Beta distribution with shape parameters $\alpha = 1.5$ and β (which can be modified, by default $\beta = 6$). The larger the β value, the lower the probability of having high fold-changes.

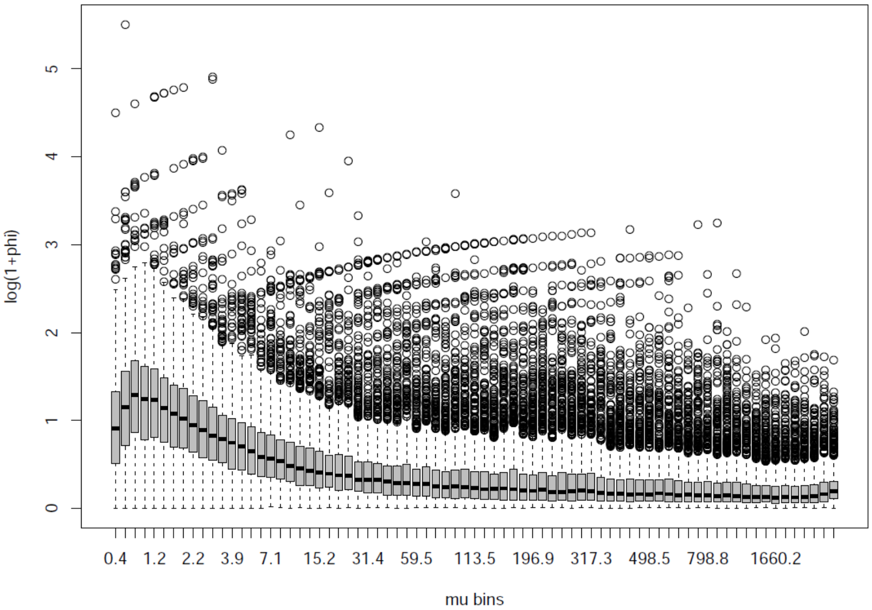


Figure S25: Distribution of ϕ values (in log-scale) from experimental data sets with high biological variability within each bin of μ values (containing approximately 1000 values each).

References

[1] McIntyre, L., Lopiano, K., Morse, A., Amin, V., Oberg, A., Young, L., and Nuzhdin, S. (June, 2011) RNA-seq: technical variability and sampling. *BMC Genomics,* **12**(1), 293+.

[2] Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Hubbard, T. J., Guigó, R., Harrow, J., Bertone, P., Consortium, R., et al. (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nature methods,*.

[3] Bullard, J., Purdom, E., Hansen, K., and Dudoit, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics,* **11**(1), 94+.

[4] Soneson, C. and Delorenzi, M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics,* **14**(1), 91.

[5] Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol,* **11**(10), R106.

[6] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* **26**(1), 139–140.

[7] Hardcastle, T. and Kelly, K. (2010) baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics,* **11**(1), 422+.

[8] Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., and Taylor, J. M. (2012) Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC genomics,* **13**(1), 484.